10.767.196 340300993

07.20.04

# 日本国特許庁
# JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年１０月１５日

出願番号
Application Number: 特願２００３－３５４５５７

[ST. 10/C]: [ＪＰ２００３－３５４５５７]

出願人
Applicant(s): 株式会社日立製作所

**CERTIFIED COPY OF PRIORITY DOCUMENT**

２００４年 ６月２１日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫

出証番号 出証特２００４－３０５３５８２

【書類名】　特許願
【整理番号】　340300993
【あて先】　特許庁長官殿
【国際特許分類】　G06F 03/06
【発明者】
　【住所又は居所】　神奈川県小田原市中里３２２番２号　株式会社日立製作所ＲＡＩＤシステム事業部内
　【氏名】　佐藤　元
【発明者】
　【住所又は居所】　神奈川県小田原市中里３２２番２号　株式会社日立製作所ＲＡＩＤシステム事業部内
　【氏名】　本間　繁雄
【特許出願人】
　【識別番号】　000005108
　【氏名又は名称】　株式会社日立製作所
【代理人】
　【識別番号】　100095371
　【弁理士】
　【氏名又は名称】　上村　輝之
【選任した代理人】
　【識別番号】　100089277
　【弁理士】
　【氏名又は名称】　宮川　長夫
【選任した代理人】
　【識別番号】　100104891
　【弁理士】
　【氏名又は名称】　中村　猛
【手数料の表示】
　【予納台帳番号】　043557
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　特許請求の範囲　1
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　0110323

【書類名】特許請求の範囲
【請求項１】
上位装置と通信可能に接続され得るディスクアレイ装置において、
　このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、
　前記上位装置とのデータ転送を制御する上位側転送制御部と、
　少なくとも、１つのパリティグループを構成する複数のデータディスクドライブと、１以上のスペアディスクドライブとを有するディスクアレイであって、前記１つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの２以上のセットに分けることができる、前記ディスクアレイと、
　前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、
　前記ディスクアレイとのデータ転送を制御する下位側転送制御部と
を備え、
　前記ディスクアレイ制御部が、
　前記データディスクドライブ毎の故障発生の可能性を推定する推定部と、
　前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて２以上のデータディスクドライブを、分割データコピーの対象として選び、前記選ばれた２以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより２以上の分割記憶領域を選び、前記選ばれた２以上の分割記憶領域は、前記パリティグループ内の前記データストライプの異なるセットにそれぞれ属し、そして、前記選ばれた２以上の分割記憶領域のデータを前記１以上のスペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する分割データコピー部と
を有するディスクアレイ装置。
【請求項２】
請求項1記載のディスクアレイ装置において、
　前記ディスクアレイ制御部が、
　前記複数のデータディスクドライブの中から、前記推定された故障発生の可能性に応じて１つのデータディスクドライブを、ダイナミックスペアリングの対象として選び、前記選ばれた１つのデータディスクドライブから、前記分割データコピー部によるコピーが未だ行われていない残りの分割記憶領域を選び、そして、前記選ばれた残りの分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御するダイナミックスペアリング部を更に有するディスクアレイ装置。
【請求項３】
請求項２記載のディスクアレイ装置において、
　前記分割データコピー部は、前記複数のデータディスクドライブの内の第１のデータディスクドライブの前記推定された故障発生の可能性が第１のレベルに達した場合、少なくとも、前記第１のデータディスクドライブと他の第２のデータディスクドライブとを、前記分割データコピーの対象として選択し、
　前記ダイナミックスペアリング部は、前記第１のデータディスクドライブの前記推定された故障発生の可能性が前記第１のレベルより高い第２のレベルに達した場合、前記第１のデータディスクドライブを、前記ダイナミックスペアリングの対象として選択するディスクアレイ装置。
【請求項４】
請求項１記載のディスクアレイ装置において、
　前記分割データコピー部は、前記複数のデータディスクドライブの内の第１のデータディスクドライブの前記推定された故障発生の可能性が所定のレベルに達した場合、前記第１のデータディスクドライブと、前記推定された故障発生の可能性が前記第１のデータディスクドライブに次いで高い第２のデータディスクドライブとを、前記分割データコピー

の対象として選択するディスクアレイ装置。
【請求項５】
請求項１記載のディスクアレイ装置において、
　前記分割データコピー部は、前記選ばれた２以上の分割記憶領域のデータをコピーする過程で、前記選ばれた２以上の分割記憶領域から同時的にデータを読み出すように前記下位側転送制御部及び前記キャッシュメモリを制御するディスクアレイ装置。
【請求項６】
請求項１記載のディスクアレイ装置において、
　前記ディスクアレイ制御部が、
　前記選ばれた２以上の分割記憶領域のデータの前記１以上のスペアディスクドライブへのコピーが開始された後、前記上位装置からの前記選ばれた２以上の分割記憶領域への新たなデータの書き込み要求を前記上位側データ転送部から受けた場合、前記上位装置から受けた前記新たなデータを、前記選ばれた２以上の分割記憶領域へ書き込むと共に、前記スペアディスクドライブにも書き込むように前記下位側転送制御部及び前記キャッシュメモリを制御するスペアデータ最新化部を更に有するディスクアレイ装置。
【請求項７】
請求項１記載のディスクアレイ装置において、
　前記分割データコピー部が、前記複数のデータディスクドライブから前記分割データコピーの対象として第１及び第２のデータディスクドライブを選び、前記第１のデータディスクドライブから前記パリティグループの内の前側のデータストライプのセットに属する第１の分割記憶領域を選び、前記第２のデータディスクドライブから前記前側のデータストライプのセットに引き続く後側のデータストライプのセットに属する第２の分割記憶領域を選び、そして、前記第１と第２の分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御するディスクアレイ装置。
【請求項８】
請求項１記載のディスクアレイ装置において、
　前記分割データコピー部により前記２以上のデータディスクドライブから選ばれる前記２以上の分割記憶領域が、実質的に等しいサイズをもつディスクアレイ装置。
【請求項９】
請求項１記載のディスクアレイ装置において、
　前記分割データコピー部により前記２以上のデータディスクドライブから選ばれる前記２以上の分割記憶領域が、前記２以上のデータディスクドライブの前記推定された故障発生の可能性に応じて異なるサイズを有するディスクアレイ装置。
【請求項１０】
請求項１記載のディスクアレイ装置において、
　前記推定部が、前記データディスクドライブ毎のエラー発生履歴を記憶し、前記記憶されたエラー発生履歴に基づいて前記データディスクドライブ毎に前記故障発生の可能性を推定するディスクアレイ装置。
【請求項１１】
上位装置と通信可能に接続され得るディスクアレイ装置であって、
　このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、
　少なくとも、１つのパリティグループを構成する複数のデータディスクドライブと、１以上のスペアディスクドライブとを有するディスクアレイであって、前記１つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの２以上のセットに分けることができる、前記ディスクアレイと、
　前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、

前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備えたものにおいて、
前記ディスクアレイ制御部が前記スペアディスクドライブを用いて前記データディスクドライブ内のデータをスペアするための方法において、
前記データディスクドライブ毎の故障発生の可能性を推定するステップと、
前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて、２以上のデータディスクドライブを、分割データコピーの対象として選ぶステップと、
前記選ばれた２以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより２以上の分割記憶領域を選ぶステップであって、前記選ばれた２以上の分割記憶領域が、前記パリティグループ内の異なるデータストライプのセットにそれぞれ属するようになったステップと、
前記選ばれた２以上の分割記憶領域のデータを前記１以上のスペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御することで分割データコピーを行うステップと
を有することを特徴とする方法。
【請求項１２】
請求項１１記載の方法において、
前記複数のデータディスクドライブの中から、前記推定された故障発生の可能性に応じて１つのデータディスクドライブを、ダイナミックスペアリングの対象として選ぶステップと、
前記選ばれた１つのデータディスクドライブから、前記スペアディスクへのコピーが未だ行われていない残りの分割記憶領域を選び、前記選ばれた残りの分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御することでダイナミックスペアリングを行うステップを更に有する方法。
【請求項１３】
請求項１２記載の方法において、
前記２以上のデータディスクドライブを選ぶステップは、前記複数のデータディスクドライブの内の第１のデータディスクドライブの前記推定された故障発生の可能性が第１のレベルに達した場合に行なわれて、そこで、少なくとも前記第１のデータディスクドライブと他の第２のデータディスクドライブとが、前記分割データコピーの対象として選択され、
前記１つのデータディスクドライブを選ぶステップは、前記第１のデータディスクドライブの前記推定された故障発生の可能性が前記第１のレベルより高い第２のレベルに達した場合に行なわれて、そこで、前記第１のデータディスクドライブが、前記ダイナミックスペアリングの対象として選択される方法。
【請求項１４】
請求項１１記載の方法において、
前記２以上のデータディスクドライブを選ぶステップは、前記複数のデータディスクドライブの内の第１のデータディスクドライブの前記推定された故障発生の可能性が所定のレベルに達した場合に行なわれて、そこで、前記第１のデータディスクドライブと、前記推定された故障発生の可能性が前記第１のデータディスクドライブに次いで高い第２のデータディスクドライブとが、前記分割データコピーの対象として選択される方法。
【請求項１５】
請求項１１記載の方法において、
前記制御するステップでは、前記選ばれた２以上の分割記憶領域から同時的にデータが読み出されように前記下位側転送制御部及び前記キャッシュメモリを制御する方法。
【請求項１６】
請求項１１記載の方法において、

前記分割データコピーを行うステップが開始された後、前記上位装置からの前記選ばれた２以上の分割記憶領域への新たなデータの書き込み要求を前記上位側データ転送部から受けた場合、前記上位装置から受けた前記新たなデータを、前記選ばれた２以上の分割記憶領域へ書き込むと共に前記スペアディスクドライブにも書き込むように、前記下位側転送制御部及び前記キャッシュメモリを制御することにより、前記スペアディスク内のデータを最新化するステップを更に有する方法。

【請求項１７】

請求項１１記載の方法において、

前記２以上のデータディスクドライブを選ぶステップでは、前記複数のデータディスクドライブから第１及び第２のデータディスクドライブが選ばれ、

前記２以上の分割記憶領域を選ぶステップでは、前記第１のデータディスクドライブから前記パリティグループ内の前側のデータストライプのセットに属する第１の分割記憶領域が選ばれ、前記第２のデータディスクドライブから前記前側のデータストライプのセットに引き続く後側のデータストライプのセットに属する第２の分割記憶領域が選ばれる方法。

【請求項１８】

請求項１１記載の方法において、

前記２以上の分割記憶領域を選ぶステップでは、実質的に等しいサイズを有する２以上の分割記憶領域が選ばれる方法。

【請求項１９】

請求項１１記載のデータスペアリング方法において、

前記２以上の分割記憶領域を選ぶステップでは、前記２以上のデータディスクドライブについて推定された故障発生の可能性に応じて、異なるサイズを有する２以上の分割記憶領域が選ばれる方法。

【請求項２０】

上位装置と通信可能に接続され得るディスクアレイ装置であって、

このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、

少なくとも、１つのパリティグループを構成する複数のデータディスクドライブと、１以上のスペアディスクドライブとを有するディスクアレイであって、前記１つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの２以上のセットに分けることができる、前記ディスクアレイと、

前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、

前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備えたものにおける、

前記スペアディスクドライブを用いて前記データディスクドライブ内のデータをスペアする動作を制御するための装置において、

前記データディスクドライブ毎の故障発生の可能性を推定する推定部と、

前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて２以上のデータディスクドライブを、分割データコピーの対象として選び、前記選ばれた２以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより２以上の分割記憶領域を選ぶものであって、前記選ばれた２以上の分割記憶領域が前記パリティグループ内の異なるデータストライプのセットにそれぞれ属するようにする分割領域選択部と、

前記選ばれた２以上の分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する分割データコピー部とを備えたデータスペアリング制御装置。

【書類名】明細書
【発明の名称】スペアディスクドライブをもつディスクアレイ装置及びデータスペアリング方法
【技術分野】
【０００１】
本発明は、RAID機能を有する複数のディスクドライブとスペアディスクドライブを持つディスクアレイ装置において、複数のディスクドライブから１台または複数台のスペアディスクドライブへ並列にデータコピーする機能を有するディスクアレイ装置およびその制御方式に関する。
【背景技術】
【０００２】
従来、ディスクアレイ装置では、RAID機能に加えてスペアディスクドライブを持ち、信頼性を向上させている。このようなディスクアレイ装置ではRAIDを構成するディスクドライブの1台に障害が発生したとき、RAIDを構成している他のディスクドライブからデータを復旧してスペアディスクドライブへコピーすること（コレクションコピー）で、RAIDの縮退動作から通常アクセス状態へ復旧する。しかし、データ復旧の際に縮退動作となるため、信頼性と性能の面で問題がある。そこで、ディスクドライブのエラー履歴を解析して故障の可能性を予測し、その可能性の高いディスクドライブのデータを、故障発生前にスペアディスクドライブへコピーすること（ダイナミックスペアリング）で、より信頼性を高めている（特許文献１、２参照）。
【０００３】
【特許文献１】特開平１１－３４５０９５号公報
【特許文献２】特開平０８－２４９１３３号公報
【発明の開示】
【発明が解決しようとする課題】
【０００４】
しかしながら、最近のディスクドライブは大容量化し、高負荷環境で使用されることも多く、そのため、上述したコレクションコピーやダイナミックスペアリングにかかる時間が増大している。そのため、次のような多重ディスク故障の発生可能性が高まっている。例えば、ダイナミックスペアリングのためのコピー動作中に、コピー元ディスクドライブで障害が頻発しREAD動作を継続できなくなる場合があり得る。この場合、当該ディスクドライブは閉塞され、このディスクドライブが属するパリティグループ（つまり、ECC（エラーコレクティングコード）グループ）についてデータ復旧動作（コレクションコピー）が行われることになる。しかしながら、このコレクションコピー動作中に、同一パリティグループ中の別のディスクドライブで、部分的なセクタ障害或いはハードディスク閉塞などの障害が発生する場合があり得る。このような多重ディスク故障が発生すると、データは喪失される。
【０００５】
従って、本発明の目的は、スペアディスクを持つRAID構成のディスクアレイ装置において、ディスクの多重故障によるデータロストの危険性を低減することにある。
【０００６】
本発明の別の目的は、高負荷時にも高速なダイナミックスペアリングが可能なディスクアレイ装置を提供することにある。
【課題を解決するための手段】
【０００７】
本発明の一つの観点に従う、上位装置と通信可能に接続され得るディスクアレイ装置は、このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、少なくとも、１つのパリティグループを構成する複数のデータディスクドライブと、１以上のスペアディスクドライブとを有するディスクアレイであって、前記１つのパリティグループは前記複数のデータディスクドライ

ブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの２以上のセットに分けることができる、前記ディスクアレイと、前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備える。そして、前記ディスクアレイ制御部が、前記データディスクドライブ毎の故障発生の可能性を推定する推定部と、前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて２以上のデータディスクドライブを、分割データコピーの対象として選び、前記選ばれた２以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより２以上の分割記憶領域を選び、前記選ばれた２以上の分割記憶領域は、前記パリティグループ内の前記データストライプの異なるセットにそれぞれ属し、そして、前記選ばれた２以上の分割記憶領域のデータを前記１以上のスペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する分割データコピー部とを有する。

【０００８】

一つの実施形態においては、前記ディスクアレイ制御部が、前記複数のデータディスクドライブの中から、前記推定された故障発生の可能性に応じて１つのデータディスクドライブを、ダイナミックスペアリングの対象として選び、前記選ばれた１つのデータディスクドライブから、前記分割データコピー部によるコピーが未だ行われていない残りの分割記憶領域を選び、そして、前記選ばれた残りの分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御するダイナミックスペアリング部を更に有する。

【０００９】

一つの実施形態においては、前記分割データコピー部は、前記複数のデータディスクドライブの内の第１のデータディスクドライブの前記推定された故障発生の可能性が第１のレベルに達した場合、少なくとも、前記第１のデータディスクドライブと他の第２のデータディスクドライブとを、前記分割データコピーの対象として選択する。前記ダイナミックスペアリング部は、前記第１のデータディスクドライブの前記推定された故障発生の可能性が前記第１のレベルより高い第２のレベルに達した場合、前記第１のデータディスクドライブを、前記ダイナミックスペアリングの対象として選択する。

【００１０】

一つの実施形態においては、前記分割データコピー部は、前記複数のデータディスクドライブの内の第１のデータディスクドライブの前記推定された故障発生の可能性が所定のレベルに達した場合、前記第１のデータディスクドライブと、前記推定された故障発生の可能性が前記第１のデータディスクドライブに次いで高い第２のデータディスクドライブとを、前記分割データコピーの対象として選択する。

【００１１】

一つの実施形態においては、前記分割データコピー部は、前記選ばれた２以上の分割記憶領域のデータをコピーする過程で、前記選ばれた２以上の分割記憶領域から同時的にデータを読み出すように前記下位側転送制御部及び前記キャッシュメモリを制御する。

【００１２】

一つの実施形態においては、前記ディスクアレイ制御部が、前記選ばれた２以上の分割記憶領域のデータの前記１以上のスペアディスクドライブへのコピーが開始された後、前記上位装置からの前記選ばれた２以上の分割記憶領域への新たなデータの書き込み要求を前記上位側データ転送部から受けた場合、前記上位装置から受けた前記新たなデータを、前記選ばれた２以上の分割記憶領域へ書き込むと共に、前記スペアディスクドライブにも書き込むように前記下位側転送制御部及び前記キャッシュメモリを制御するスペアデータ最新化部を更に有する。

【００１３】

一つの実施形態においては、前記分割データコピー部が、前記複数のデータディスクドライブから前記分割データコピーの対象として第１及び第２のデータディスクドライブを

選び、前記第１のデータディスクドライブから前記パリティグループの内の前側のデータストライプのセットに属する第１の分割記憶領域を選び、前記第２のデータディスクドライブから前記前側のデータストライプのセットに引き続く後側のデータストライプのセットに属する第２の分割記憶領域を選び、そして、前記第１と第２の分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する。

【００１４】

一つの実施形態においては、前記２以上のデータディスクドライブから選ばれる前記２以上の分割記憶領域が、実質的に等しいサイズをもつ。

【００１５】

一つの実施形態においては、前記２以上のデータディスクドライブから選ばれる前記２以上の分割記憶領域が、前記２以上のデータディスクドライブの前記推定された故障発生の可能性に応じて異なるサイズを有する。

【００１６】

一つの実施形態においては、前記推定部が、前記データディスクドライブ毎のエラー発生履歴を記憶し、前記記憶されたエラー発生履歴に基づいて前記データディスクドライブ毎に前記故障発生の可能性を推定する。

【００１７】

本発明の別の観点に従うデータスペアリング方法は、上位装置と通信可能に接続され得るディスクアレイ装置であって、このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、少なくとも、１つのパリティグループを構成する複数のデータディスクドライブと、１以上のスペアディスクドライブとを有するディスクアレイであって、前記１つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの２以上のセットに分けることができる、前記ディスクアレイと、前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備えたものに適用される。この方法は、前記データディスクドライブ毎の故障発生の可能性を推定するステップと、前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて、２以上のデータディスクドライブを、分割データコピーの対象として選ぶステップと、前記選ばれた２以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより２以上の分割記憶領域を選ぶステップであって、前記選ばれた２以上の分割記憶領域が、前記パリティグループ内の異なるデータストライプのセットにそれぞれ属するようになったステップと、前記選ばれた２以上の分割記憶領域のデータを前記１以上のスペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御することで分割データコピーを行うステップとを有する。

【００１８】

この方法の一つの実施形態は、前記複数のデータディスクドライブの中から、前記推定された故障発生の可能性に応じて１つのデータディスクドライブを、ダイナミックスペアリングの対象として選ぶステップと、前記選ばれた１つのデータディスクドライブから、前記スペアディスクへのコピーが未だ行われていない残りの分割記憶領域を選び、前記選ばれた残りの分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御することでダイナミックスペアリングを行うステップを更に有する。

【００１９】

一つの実施形態においては、前記２以上のデータディスクドライブを選ぶステップは、前記複数のデータディスクドライブの内の第１のデータディスクドライブの前記推定された故障発生の可能性が第１のレベルに達した場合に行なわれて、そこで、少なくとも前記第１のデータディスクドライブと他の第２のデータディスクドライブとが、前記分割デー

タコピーの対象として選択される。また、前記１つのデータディスクドライブを選ぶステップは、前記第１のデータディスクドライブの前記推定された故障発生の可能性が前記第１のレベルより高い第２のレベルに達した場合に行なわれて、そこで、前記第１のデータディスクドライブが、前記ダイナミックスペアリングの対象として選択される。

【００２０】

本発明のまた別の観点に従う、データスペアリング制御装置は、上位装置と通信可能に接続され得るディスクアレイ装置であって、このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、少なくとも、１つのパリティグループを構成する複数のデータディスクドライブと、１以上のスペアディスクドライブとを有するディスクアレイであって、前記１つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの２以上のセットに分けることができる、前記ディスクアレイと、前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備えたものに適用される。このデータスペアシング制御装置は、前記データディスクドライブ毎の故障発生の可能性を推定する推定部と、前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて２以上のデータディスクドライブを、分割データコピーの対象として選び、前記選ばれた２以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより２以上の分割記憶領域を選ぶものであって、前記選ばれた２以上の分割記憶領域が前記パリティグループ内の異なるデータストライプのセットにそれぞれ属するようにする分割領域選択部と、前記選ばれた２以上の分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する分割データコピー部とを備える。

【発明の効果】

【００２１】

本発明に従うディスクアレイ装置によれば、或るパリティグループを形成する複数のデータディスクドライブの中から、それぞれの故障発生の可能性に応じて２以上のデータディスクドライブが選ばれ、それらのデータディスクドライブの各々の記憶領域が分割されて、そこから、異なるデータストライプのセットにそれぞれ属する２以上の分割記憶領域が選ばれ、それら２以上の分割記憶領域のデータが、それらのデータディスクドライブが故障する前に、スペアディスクドライブにコピーされる。このような分割データコピー処理が故障発生の可能性がある程度に高くなったデータディスクドライブについて行われることより、それらのデータディスクドライブの故障可能性が後に上昇した場合、分割データコピー処理が未だ行われていない残りの分割記憶領域についてのみ、例えばダイナミックスペアリング処理のようなデータ保全のための処理を行えばよいことになる。その結果、データ保全に要する処理時間が短縮できる。また、分割データコピー処理においては、２以上のディスクドライブの分割記憶領域から並列的にデータをREADするようにすれば、ディスクドライブのREAD時間も短縮される。結果として、高負荷環境においても多重ディスク故障によるデータロストの危険性が低減する。

【発明を実施するための最良の形態】

【００２２】

以下、本発明に従うディスクアレイ装置の一実施形態を説明する。

【００２３】

図１は、本発明の一実施形態に係るディスクアレイ装置を使用したコンピュータシステムの概略構成、及びこのディスクアレイ装置の内部の構成を示す。

【００２４】

図１に示すように、ディスクアレイ装置１００は、ホストコンピュータのような上位装置１１０と接続されており、複数のハードディスクドライブ（ドライブ番号＃０～＃N）３００～３３０と１又は２以上のコントローラ２００、２１０を備えている。また、この

ディスクアレイ装置１００は、その操作・設定を行うための制御コンソール１２０を備え、てコントローラ２００、２１０がLANインタフェース２０１、２１１を介して制御コンソール１２０にそれぞれ接続されている。

【００２５】

コントローラ２００、２１０はそれぞれ、上位装置１００とのデータ転送を制御する上位側転送制御部２０２、２１２、ディスクアレイ装置１００全体の制御を行うディスクアレイ制御部２０３、２１３、上位装置１００及びディスクドライブ３００～３３０のデータ転送等に用いられるキャッシュメモリ２３０、及び、ディスクドライブ３００～３３０のデータ転送を制御する下位側転送制御部２０４、２１４を備える。コントローラ２００は、ポート４００～４３０をそれぞれ介してディスクドライブ３００～３３０にアクセスする。コントローラ２１０は、ポート４４０～４７０をそれぞれ介してディスクドライブ３００～３３０にアクセスする。本実施形態の以下の説明では、複数のディスクドライブ３００～３３０の物理的な並びは「ディスクアレイ」と呼ばれ、RAID技術による冗長化された構成の論理ボリュームを「LU（論理ユニット）」と呼ばれる。

【００２６】

次に、図２及び図３を用いて、本発明のディスクアレイ装置１００が用いるディスクアレイの状態を管理するためのデータ構造について説明する。図２は、ディスクドライブ資源情報テーブル５００を例示する模式図である。図３はスペアディスクドライブ資源情報テーブル６００を例示する模式図である。

【００２７】

図２に示されたディスクドライブ資源情報テーブル５００は、コントローラ２００、２１０のディスクアレイ制御部２０３、２１３の各々に存在する。ディスクドライブ資源情報テーブル５００は、全てのディスクドライブ３００～３３０の使用状況や容量などを管理するためのテーブルであり、全てのディスクドライブ３００～３３０にそれぞれ対応する複数のディスクドライブ管理領域５１０を有する。各ディスクドライブ管理領域５１０に割り当てられたROW（行）番号（ROW#0、ROW#1、…）及びCOL（列）番号（COL#0、COL#1、…）は、対応するディスクドライブのディスクアレイ装置１００内での位置を示している。

【００２８】

各ディスクドライブ管理領域５１０は、対応するディスクドライブについて、例えば、使用状態５２０、全容量５３０、スペアリング済み容量（又はアドレス）５４０、及びエラーカウンタ５５０などの情報を有する。使用状態５２０には、対応するディスクドライブについて実装状況及び使用状況を示す値、例えば、未実装、実装（データディスクドライブとして使用されるか、スペアディスクドライブとして使用されるかも区別する）、及び未使用の４つの値が選択的に入る。全容量５３０は、対応するディスクドライブが実装されている場合に、そのディスクドライブの全記憶容量を表す値が入る。スペアリング済容量（又はアドレス）５４０は、対応するディスクドライブの故障発生可能性が高まった場合にスペアディスクドライブへのデータコピー時に用いられるフィールドであり、そのディスクドライブの全容量の内で現在時点でスペアディスクドライブへのデータコピーが行われた容量（又はデータコピーが行われた記憶領域のアドレス（例えば、論理ブロック番号など））を示す値が入る。エラーカウンタ５５０には、対応するディスクドライブにおけるアクセスエラーの発生回数を示す値が入り、この値はディスクドライブでアクセスエラーが発生する都度にカウントアップ（又はカウントダウン）される。エラーカウンタ５５０の値は、後述するエラーレート（エラーの発生頻度）、換言すれば故障発生の可能性の高さを表すものとして用いられる。

【００２９】

図３に示されたスペアディスクドライブ資源情報テーブル６００は、コントローラ２００、２１０のディスクアレイ制御部２０３、２１３の各々に存在する。スペアディスクドライブ資源情報テーブル６００は、ディスクドライブ３００～３３０の内、スペアディスクドライブとして使用されるもの状態を管理するためのテーブルであり、全てのスペアデ

ィスクドライブにそれぞれ対応する複数のスペアディスクドライブ情報領域６１０を有する。

【００３０】

各スペアディスクドライブ情報６１０は、対応するスペアディスクドライブについての詳細情報、例えば、使用状態６２０、スペアＲＯＷ６３０、スペアＣＯＬ６４０、使用領域情報［0］６５０～［1］６６０などの情報を含む。使用状態６２０は、対応するスペアディスクドライブの使用状態を示す値、例えば、未使用、一部使用、及び全使用の３つの値を選択的に持つ。使用状態６２０は、或るデータディスクドライブのデータのスペアリングを行う場合に、そのスペアディスクドライブの使用可否判断を行うために使用される。スペアROW６３０及びスペアCOL６４０には、対応するスペアディスクドライブのディスクアレイ装置１００内での位置を示すROW番号及びCOL番号が入る。使用領域情報［0］６５０～［m］６７０は、対応するスペアディスクドライブに含まれるm個の記憶領域（［0］～［m］）（mは、そのスペアディスクドライブの分割数であり、後述する図４に示す例ではｍ＝２である）にそれぞれ対応して設けられ、それらm個の記憶領域がそれぞれどのディスクドライブのデータ復元用に使用されたかを管理するための情報である。各使用領域情報は、使用中フラグ６７０、開始アドレス６８０、終了アドレス６９０、コピー元ROW７００、及びコピー元COL７１０などの情報を有する。使用中フラグ６７０は、対応する記憶領域が現在使用中であるか否かを示す。開始アドレス６８０及び終了アドレス６９０は、対応する記憶領域の開始アドレスと終了アドレスを示す。コピー元ROW７００及びコピー元COL７１０には、対応する記憶領域にコピーされたデータの元々記憶されてたデータディスクドライブのディスクアレイ装置１００内での位置を示すROW番号とCOL番号が入る。

【００３１】

ディスクアレイ装置１００のディスクアレイ制御部２０３、２１３（図１参照）は、上述したディスクドライブ資源情報テーブル５００及びスペアディスクドライブ資源情報テーブル６００を用いて、ディスクドライブ３００～３３０の状態を把握し、そして、ディスクドライブの故障からデータを保護するために、以下に説明する制御動作を行う。

【００３２】

以下の説明では、一例として、各ディスクドライブのエラーレート（アクセスエラーの発生頻度）が、各ディスクドライブの故障発生の可能性を示す３段階のレベルに分類される。すなわち、「レベル１」は故障発生確率が低いことを、「レベル２」は故障発生確率が中程度であることを、そして、「レベル３」は故障発生確率が高いことを意味する。ディスクアレイ制御部２０３、２１３は、或るディスクドライブのエラーレートが例えばレベル２になったとき、そのディスクドライブについて本発明の原理に従う「スペアディスクへの分割データコピー」を行なって、故障発生に備える。その後、そのディスクドライブでさらに多くのエラーが発生して、エラーレートが例えばレベル３になると、ディスクアレイ制御部２０３、２１３は、そのディスクドライブについてダイナミックスペアリングを実施し、そして、ダイナミックスペアリングの完了後、そのディスクドライブを閉塞する。

【００３３】

このようなディスクドライブの故障からデータを保護するための制御について、以下により具体的に説明する。以下の説明では、RAID－5技術に従い3D+1P形式のパリティグループを形成するように構成されたつディスクアレイの場合を例にとるが、RAID技術に従がう他の構成のディスクアレイにおいても、本発明の原理に従う制御が適用できることは当業者は容易に理解できるはずである。

【００３４】

図４は、3D+1P形式のパリティグループを形成するディスクアレイの構成例を示す。

【００３５】

図４に示す例では、3D+1P形式の１つのパリティグループを形成する４つのデータディスクドライブ（DDD#0～DDD#3）８００～８３０からなるディスクアレイが存在する。この

ディスクアレイ８００～８３０内に、例えば２つの論理ユニットLU0、LU1が定義されている。一方の論理ユニットLU0は論理ブロックBlock0～Block17から構成され、他方の論理ユニットLU1は論理ブロックBlock18～BlockXで構成されている（１論理ブロックのサイズは例えば６４Kバイトである）。これら４つの異なるデータディスクドライブ８００～８３０にそれぞれ含まれる図中横方向一列に並ぶ４つの論理ブロックが、3D+1P形式の一つのデータストライプを形成する。ここで、各データストライプ内のP(Q-R)と記された論理ブロックは、同じデータストライプ内の他の論理ブロックBlockQ～BlockRのデータのパリティデータを記憶したブロックであることを意味する。例えば、第１のデータストライプは、３つのデータ論理ブロックBlock0～Block2と１つのパリティ論理ブロックP(0-2)から構成され、また、第２のデータストライプは、３つのデータブロックBlock3～Block5と１つのパリティブロックP(3-5)から構成される。このようにして、１つのパリティグループを形成するディスクアレイ８００～８３０内には、多数のデータストライプが存在する。さらに、このディスクアレイ８００～８３０の他に、少なくとも１つのスペアディスクドライブ（SDD#A)が存在する。

【００３６】

図５は、図４に例示したような構成のディスクアレイ８００～８３０に適用される、ディスクアレイ制御部２０３、２１３（図１参照）による、データディスクドライブの故障からデータを保護するための制御の全体的な流れを示す。

【００３７】

図５に示すように、ディスクアレイ制御部２０３、２１３は、ディスクアレイ８００～８３０内に、そのエラーレートがレベル２に達したディスクドライブが存在する否かチェックする（ステップ９００）。このチェックは、図２に示されたディスクドライブ資源情報テーブル５００内から、そのエラーカウンタ５５０の値がレベル２に相当する所定の閾値に達したディスクドライブ管理領域５１０を探すことで行うことができる。そのようなディスクドライブ管理領域５１０が見つかった場合、そのディスクドライブ管理領域５１０に対応するROW番号とCOL番号により決まる位置にあるディスクドライブが、エラーレートがレベル２に達したディスクドライブとして検出される。そのようなディスクドライブが検出された場合、ディスクアレイ制御部２０３、２１３は、図３に示されたスペアディスクドライブ資源情報テーブル６００を参照し、使用可能なスペアディスクドライブの有無を判定する（ステップ９１０）。使用可能なスペアディスクが有る場合、ディスクアレイ制御部２０３、２１３は、下位側転送制御部２０４、２１４及びキャッシュメモリ２２０を制御して、上記検出されたディスクドライブについて、分割データコピー処理を行う（ステップ９２０）。分割データコピー処理の具体的方法については後に説明する。他方、使用可能なスペアディスクがない場合には、ディスクアレイ制御部２０３、２１３は、制御コンソール１１０に対して、ディスクドライブのエラーレート上昇を知らせる警告を発し、オペレータの注意を促す（ステップ９５０）。

【００３８】

また、ディスクアレイ制御部２０３、２１３は、ディスクアレイ８００～８３０内に、そのエラーレートがレベル３に達したディスクドライブが存在する否かチェックする（ステップ９３０）。このチェックは、図２に示されたディスクドライブ資源情報テーブル５００内から、そのエラーカウンタ５５０の値がレベル３に相当する所定の閾値に達したディスクドライブ管理領域５１０を探すことで行うことができる。そのようなディスクドライブ管理領域５１０が見つかった場合、そのディスクドライブ管理領域５１０に対応するROW番号とCOL番号により決まる位置にあるディスクドライブが、エラーレートがレベル３に達したディスクドライブとして検出される。そのようなディスクドライブが検出された場合、ディスクアレイ制御部２０３、２１３は、下位側転送制御部２０４、２１４及びキャッシュメモリ２２０を制御して、その検出されたディスクドライブについてダイナミックスペアリングを実行する（ステップ９４０）。ここで、上述の説明から明らかなように、ダイナミックスペアリングの対象となったディスクドライブについては、通常、既にステップ９２０の分割データコピーが行われている。後述する分割データコピーの具体的な

処理から分かるように、分割データコピーによって、そのディスクドライブ内の全記憶領域の内の半分の記憶領域のデータがスペアディスクドライブに既にコピーされ終わっている。そのため、ダイナミックスペアリングでは、残り半分の記憶領域のデータをスペアディスクドライブへコピーするだけでよい。よって、ダイナミックスペアリングにかかる時間は、従来よりも半減する。ダイナミックスペアリングの完了後、そのディスクドライブは閉塞され、正常なディスクドライブへの交換作業が行われる。図５には示されていないが、閉塞されたディスクドライブの交換完了後、そのディスクドライブにスペアディスクドライブからデータのコピーバック（データ復元）が行われる。このデータ復元が完了した後、そのスペアディスクドライブは待機状態に設定され、次の分割コピー処理及びダイナミックスペアリングの機会に利用される。

【００３９】

次に、図５中のステップ９２０の分割データコピー処理の具体的方法について説明する。

【００４０】

図４には、既に説明したディスクアレイ８００～８３０の構成例と共に、このディスクアレイ８００～８３０において実行される分割データコピー処理の具体的な態様の一例が示されている。図４を参照して、この分割データコピー処理の一例を説明する。

【００４１】

図４に示されたディスクアレイ８００～８３０において、例えば、一つのデータディスクドライブ（DDD#0）８００のエラーレートがレベル２に達したとする。このとき、他のデータディスクドライブ（DDD#1～#3）８１０～８３０のエラーレートはレベル１であったとする。ただし、レベル１のデータディスクドライブ（DDD#1～#3）８１０～８３０の中で、データディスクドライブ（DDD#1）８１０のエラーカウンタの値が最も高かったとする。また、このとき、使用可能なスペアディスク（SDD#A）８４０が存在していたとする。

【００４２】

このような状況では、レベル２に達したディスクドライブ（DDD#0）８００について、それが故障してREAD動作が不可能になる前に、分割データコピー処理が実行される。この分割データコピー処理では、ディスクドライブ（DDD#0）８００の全記憶領域の中から、図中上半分（前半）の記憶領域（論理ブロックBlock0～Block12）８５０が選ばれ、そして、選ばれた上半分記憶領域８５０のデータを、スペアディスクドライブ（SDD#A）８４０の対応する上半分記憶領域８８０へコピーする処理が実行される。すなわち、下位側転送制御部２０４又は２１４の制御により、上半分記憶領域８５０のデータがキャッシュメモリ２２０に読み出され、そして、キャッシュメモリ２２０からスペアディスクドライブ（SDD#A）８４０の対応する上半分記憶領域８８０へ書き込まれる。これと並行して、このディスクドライブ（DDD#0）８００に次いで高いエラーカウント値をもつ（つまり、次に故障発生の可能性の高い）別の一つのディスクドライブ（DDD#1）８１０が選ばれ、選ばれたディスクドライブ（DDD#1）８１０の中から、上述したディスクドライブ（DDD#0）８００内の分割コピー範囲（つまり、上半分記憶領域８５０）が属する上半分（前半）のデータストライプのセットに引き続く下半分（後半）のデータストライプのセットに該当する下半分（後半）の記録場所領域、（Block16、P7、Block21、…）８７０が選ばれ、そして、この選ばれた下半分記録場所領域（Block16、P7、Block21、…）８７０のデータを、スペアディスクドライブ（SDD#A）８４０の対応する半分記憶領域８９０へコピーする処理が実行される。すなわち、下位側転送制御部２０４又は２１４の制御により、下半分記録場所領域８７０のデータがキャッシュメモリ２２０に読み出され、そして、キャッシュメモリ２２０からスペアディスクドライブ（SDD#A）８４０の対応する半分記憶領域８９０へ書き込まれる。

【００４３】

要するに、分割データコピー処理では、１つのパリティグループを形成する複数のディスクアレイ８００～８３０の中から、ある程度高い故障発生可能性をもつ２つのディスク

ドライブ（DDD #0、#1）８００、８１０が選ばれ、選ばれた２つのディスクドライブ（DDD #0、#1）８００、８１０から、異なるデータストライプのセットにそれぞれ属する上半分（前半）と下半分（後半）の記憶領域８５０、８７０がそれぞれ選ばれ、そして、選ばれた上半分（前半）と下半分（後半）の記憶領域８５０、８７０から同時にデータがREADされて、スペアディスク８４０にWriteされる。このとき、同時に２つのディスクドライブ（DDD#0、#1）８００、８１０からREADされたデータは、ディスクアレイ制御部２０３、２１３によってスケジューリングされ、スペアディスクドライブ８４０へWRITEされる。この２つのディスクドライブからの同時データREADによって、データREADにかかる時間が半減する。これによるコピー時間の短縮は特に高負荷環境においては有利である。さらに、スペアディスクドライブ８４０が複数のアクセスポートを持つ場合は、スペアディスクドライブ８４０への多重データWRITEを行うことができるから、分割データコピー処理はより高速化される。

【００４４】

なお、分割データコピー処理の実行中、コピー対象のディスクドライブ（DDD#0、#1）８００、８１０内の既にコピーが完了した論理ブロックに対して新たなデータのWRITE要求が上位装置１１０などから上位側転送制御部２０２、２１３を通じて入力された場合には、ディスクアレイ制御部２０３、２１３は、下位側転送制御部２０４とキャッシュメモリ２２０を制御して、Write対象のディスクドライブ（DDD#0又は#1）８００又は８１０とスペアディスクドライブ（SDD#A）８４０の双方へ当該データを２重にWriteする。これにより、スペアディスクドライブ（SDD#A）８４０内のデータは常に最新状態に保たれ、コピー元のディスクドライブ（DDD#0、#1）８００、８１０の故障発生に備えることができる。

【００４５】

図６は、分割データコピー処理の具体的な流れを示す。

【００４６】

図６に示すように、分割データコピー処理が開始されると、エラーレートがレベル２に達したディスクドライブと、このディスクドライブに次いでエラーレート（エラーカウンタ値）が高いディスクドライブから、分割データコピーの対象である２つの記憶領域が選択される（ステップ９６０）。図４の例では、２つのディスクドライブ（DDD#0又は#1）８００又は８１０から、半分記憶領域８５０と８６０がそれぞれ選ばれるのである。ここで、重要な点は、選ばれた２つの記憶領域８５０と８６０は、それが属するデータストライプのセットにおいて重複しないことである。換言すれば、コピーの対象として選ばれた全ての論理ブロックは、それぞれ異なるデータストライプのセットに属するということである。これにより、出来る限り多くのデータが保護される可能性が高まるのである。

【００４７】

この後、上述した２つのディスクドライブについて並行して、それぞれのコピー対象領域内の全ての論理ブロックからデータが逐次にReadされ、Readされたデータが所定のスケジュールに従ってスペアディスクドライブの対応する論理ブロックにWriteされる（一方のドライブについてはステップ９７０及びステップ９９０～１０１０のループ、他方のドライブについてはステップ９７０及びステップ１０２０～１０５０のループ）。

【００４８】

図７は、エラーレートがレベル３に達したディスクドライブについて行われるダイナミックスペアリング処理の具体的な流れを示す。

【００４９】

図７に示すように、ダイナミックスペアリング処理が開始されると、この処理の対象となるディスクドライブの中で既にスペアリング（スペアディスクへのコピー）が完了している記憶領域がどこであるかが、ディスクアレイ制御部２０３又は２１３によって確認される（ステップ１１００）。この確認は、図２に示したディスクドライブ資源情報テーブル５００内のそのディスクドライブに対応するディスクドライブ管理領域５１０のスペアリング済容量（又はアドレス）５４０を参照することで行うことができる。通常、そのデ

ィスクドライブについては分割データコピー処理が既に行われているから、そのディスクドライブの中の半分の記憶領域は既にスペアリングが完了していることになる。その後、ディスクアレイ制御部２０３又は２１３の制御により、そのディスクドライブの中の未だスペアリングが完了していない記憶領域のデータが、ブロック番号の順で逐次に、下位側転送制御部２０４又は２１４を通じてキャッシュメモリ２２０に読み出され、キャッシュメモリ２２０から下位側転送制御部２０４又は２１４を通じてスペアディスクドライブの対応する記憶領域にコピーされる（ステップ１１１０～１１４０のループ）。

【００５０】

なお、ダイナミックスペアリングが完了する前に、ダイナミックスペアリングの対象のディスクドライブが故障してデータのREADができなくなった場合は、Read対象のデータと同じパリティグループを構成する他のディスクドライブ内のデータを用いたコレクションコピー動作により、そのRead対象のデータが復元されてスペアディスクドライブへコピーされることになる。このコレクションコピー動作が完了する前に、上述した他のディスクドライブも故障してデータのREADができなくなった場合は、従来技術によれば、これら故障したディスクドライブに関わるサブシステムはシステムダウン状態に陥らざるを得ない。しかしながら、本実施形態では、その故障した他のディスクドライブについて既に分割データコピーが行われていることが多いため、上記のような二重故障が発生しても、スペアディスクドライブ内のデータを用いて、コレクションコピーの対象のデータを復元することができる。

【００５１】

次に、図８を参照して、3D＋1P形式のディスクアレイを構成する４つのデータディクスドライブに対して２つのスペアディスクドライブが用意されている場合における、上述したデータ保護のための制御動作の具体例を説明する。

【００５２】

図８Aに示すように、3D+1のディスクアレイを構成する４つのデータディクスドライブDDD#0～#3において、ディスクドライブDDD#0のエラーレートがレベル２に達し、他のディスクドライブDDD#1～#3のエラーレートがレベル１（ただし、#2>#3>#1）であった場合、レベル２のエラーレートをもつディスクドライブDDD#0と、次に高いエラーレートをもつディスクドライブDDD#2について、スペアディスクドライブSDD#Aへの分割データコピー処理が実行される。すなわち、ディスクドライブDDD#0の上半分（前半）のデータ#0_UHとディスクドライブDDD#2の下半分（後半）のデータ#2_LHが、スペアディスクドライブSDD#Aへコピーされる。

【００５３】

その後、図８Bに示すように、ディスクドライブDDD#0のエラーレートが上昇してレベル３に達した場合、このディスクドライブDDD#0がREADできなくなる前に、このディスクドライブDDD#0についてスペアディスクドライブSDD#Aへダイナミックスペアリング処理が実行される。このダイナミックスペアリング処理では、ディスクドライブDDD#0の未だスペアリングされてない下半分（後半）のデータ#0_LHをスペアディスクドライブSDD#Aにコピーするだけでよい。このとき、スペアディスクドライブSDD#A内にあったディスクドライブDDD#2の下半分（後半）のデータ#2_LHは消去されることになる。このとき、ディスクドライブDDD#2のエラーレートもレベル２に達した場合、分割データコピーが実行されて、ディスクドライブDDD#2の上半分（前半）のデータ#2_UHと次にエラーレートの高いディスクドライブDDD#3の下半分（後半）のデータ#3_LHが、もう一つのスペアディスクドライブSDD#Bへコピーされる。

【００５４】

ディスクドライブDDD#0のダイナミックスペアリング処理の完了後、ディスクドライブDDD#0は閉塞され、別の正常なディスクドライブに交換される。分割データコピー処理所路ディスクドライブDDD#2、DDD#3は、分割データコピー処理が完了した後もエラーレートがレベル３に達しない限り、通常通り使用される。ただし、ディスクドライブDDD#2又はDDD#3へのデータのWRITEが行われるときには、ディスクドライブDDD#2又はDDD#3と共にスペ

アディスクドライブSDD#Bへも同じデータが２重にWriteされ、それにより、スペアディスクドライブSDD#B内のデータが常に最新に保たれる。

【００５５】

ディスクドライブDDD#0交換が完了した後、図８Cに示すように、スペアディスクドライブSDD#Aからディスクドライブ DDD#0へデータコピーが行われて、ディスクドライブDDD#0のデータが回復される。この回復処理が完了した後、スペアディスクドライブSDD#Aは待機状態となる。この段階で、スペアディスクドライブSDD#Bのデータは引き続き最新の状態に維持され、ディスクドライブDDD#2、DDD#3の将来のエラーレート上昇に備える。

【００５６】

図９は、図８に示したデータコピー動作の変形例を示す。

【００５７】

上述した図８のデータコピ動作では、図８Aに示された分割データコピー処理において、２つのディスクドライブDDD#0、DDD#2のデータ#0_UH、#2_LHが、同一のスペアディスクドライブSDD#Aへコピーされる。しかし、変形例として、図９Aに示すように、２つのディスクドライブDDD#0、DDD#2のデータ#0_UH、#2_LHをそれぞれ異なるスペアディスクドライブSDD#A、SDD＃Bにコピーすることもできる。そのようにすると、その後に、図９Bに示すようにディスクドライブDDD#0のエラーレートがレベル３に達してダイナミックスペアリング処理を行う場合、既にスペアリングされたディスクドライブDDD#2のデータ#2_LHを消去することなしに、ディスクドライブDDD#0の残りデータ#0_LHをスペアディスクドライブSDD#Aにコピーすることができる。その結果、ディスクドライブDDD#2のエラーレートがレベル２に達して分割データコピー処理を行う場合、ディスクドライブDDD#2のスペアリング済みデータ#2_LHのコピーを省略して、次に高いエラーレートをもつディスクドライブDDD#3のデータ#3_UHだけをコピーすることができる。

【００５８】

図１０は、図８に示したデータコピー動作のまた別の変形例を示す。

【００５９】

図１０Aに示した分割データコピー処理の動作は、図８Aに示したものと同じである。しかし、図１０Bに示したディスクドライブDDD#0のダイナミックスペアリング処理においては、ディスクドライブDDD#0の下半分（後半）のデータ#0_LHを、上半分（前半）のデータ#0_UHとは別のスペアディスクドライブSDD＃Bにコピーする。これにより、既にスペアリングされたディスクドライブDDD#2のデータ#2_LHを消去しなくてよい。その結果、図９の動作例と同様に、ディスクドライブDDD#2のエラーレートがレベル２に達して分割データコピー処理を行う場合、ディスクドライブDDD#2のスペアリング済みデータ#2_LHのコピーを省略して、次に高いエラーレートをもつディスクドライブDDD#3のデータ#3_UHだけをコピーすることができる。

【００６０】

図１１は、図４に示した分割データコピー処理の動作の変形例を示す。

【００６１】

図４に示した分割データコピー処理では、２つのディスクドライブDDD#0、DDD#1から実質的に均等なサイズの（つまり、半分サイズの）記憶領域８０５、８０７をそれぞれ選んで、それらのデータのスペアリングを行っている。変形例として、図１１に示すように、２つのディスクドライブDDD#0、DDD#1から不均等なサイズの記憶領域１２００、１２１０をそれぞれ選んで、それらのデータのスペアリングを行うこともできる。図１１の例では、エラーレート（故障発生の可能性）のより高いディスクドライブDDD#0からより大きいサイズの記憶領域１２００をコピーし、エラーレートのより低いディスクドライブDDD#1からより小さいサイズの記憶領域１２１０をコピーしている。このような不均等分割データコピー処理によれば、図４に示したような均等分割データコピー処理に比べて、分割データコピーに要する時間の短縮効果は劣るが、その分、後に行われるダイナミックスペアリング処理（エラーレートのより高いディスクドライブDDD#0について行われる可能性が高い）において処理時間が短縮できる効果が期待できる。

【００６２】
図１２は、図４に示した分割データコピー処理の動作の別の変形例を示す。
【００６３】
図４に示した分割データコピー処理では、２つのディスクドライブDDD#0、DDD#1が処理対象として選ばれる。変形例として、図１２に示すように、３つ又はそれより多くのディスクドライブDDD#0、DDD#1、DDD#2を分割データコピー処理対象として選んでもよい。図１２の例では、エラーレートの高い順に３つのディスクドライブDDD#0、DDD#1、DDD#2が選ばれ、それらからパリティグループが重ならない３つの記憶領域１３００、１３１０、１３２０がそれぞれ選ばれてコピーされる。
【００６４】
以上、本発明の実施形態を説明したが、この実施形態は本発明の説明のための例示にすぎず、本発明の範囲をこの実施形態にのみ限定する趣旨ではない。本発明は、その要旨を逸脱することなく、その他の様々な態様でも実施することができる。
【００６５】
例えば、上述した実施形態では、エラーレートを３段階のレベルに分類して、レベル２のドライブについて分割データコピーを行い、レベル３のドライブについてダイナミックスペアリングを行う。しかし、エラーレートを３段階より多くのレベルに分類して、中間的な複数のレベル（例えば、４段階のレベルのうちのレベル２と３）のドライブについて、レベルに応じた分割データコピーを行う（例えば、レベルが高いほど、スペアリングされるデータサイズが増える）ようにすることもできる。
【００６６】
また、上述の実施形態では、「RAID-5」技術に従う３D+1P形式のパリティグループを構成するディスクアレイを例にとり説明したが、「RAID-1」～「RAID-5」のいすれかに従う他の形式のパリティグループを構成するディスクアレイにも本発明が適用可能なことは明らかである。例えば、３D+1P形式のパリティグループであっても、パリティデータの配置などにおいて、図１３に示す例やその他のバリエーションが存在するが、それらバリエーションのいずれにも本発明は適用可能である。
【００６７】
また、上記実施形態では、分割データコピー及びダインミックスペアリングなどの動作を、ディスクアレイ装置が、上位装置などの外部装置からの指令によらずに自立的に制御しているが、外部装置からの指令によりそのような制御を行えるようにしてもよい。
【図面の簡単な説明】
【００６８】
【図１】本発明の一実施形態にかかるディスクアレイ装置を使用したコンピュータシステムの概略構成、及びこのディスクアレイ装置の内部の構成を示したブロック図。
【図２】ディスクドライブ資源情報テーブルの構造を示す模式図。
【図３】スペアディスクドライブ資源情報テーブルの構造を示す模式図。
【図４】3D+1P形式のパリティグループを形成するディスクアレイの構成と、スペアディスクドライブへの分割データコピーの様子を示した模式図。
【図５】ディスクアレイ制御部２０３、２１３が行うディスクアレイのデータ保護のための制御動作の流れを示すフローチャート。
【図６】分割データコピー処理の具体的な流れを示すフローチャート。
【図７】ダイナミックスペアリング処理の具体的な流れを示すフローチャート。
【図８】3D＋1P形式のディスクアレイを構成する４つのデータディクスドライブに対して２つのスペアディスクドライブが用意されている場合における、データ保護のための制御動作の具体例を示す模式図。
【図９】図８に示したデータコピー動作の変形例を示す模式図。
【図１０】図８に示したデータコピー動作の別の変形例を示す模式図。
【図１１】分割データコピー処理の変形例を示す模式図。
【図１２】分割データコピー処理の別の変形例を示す模式図。

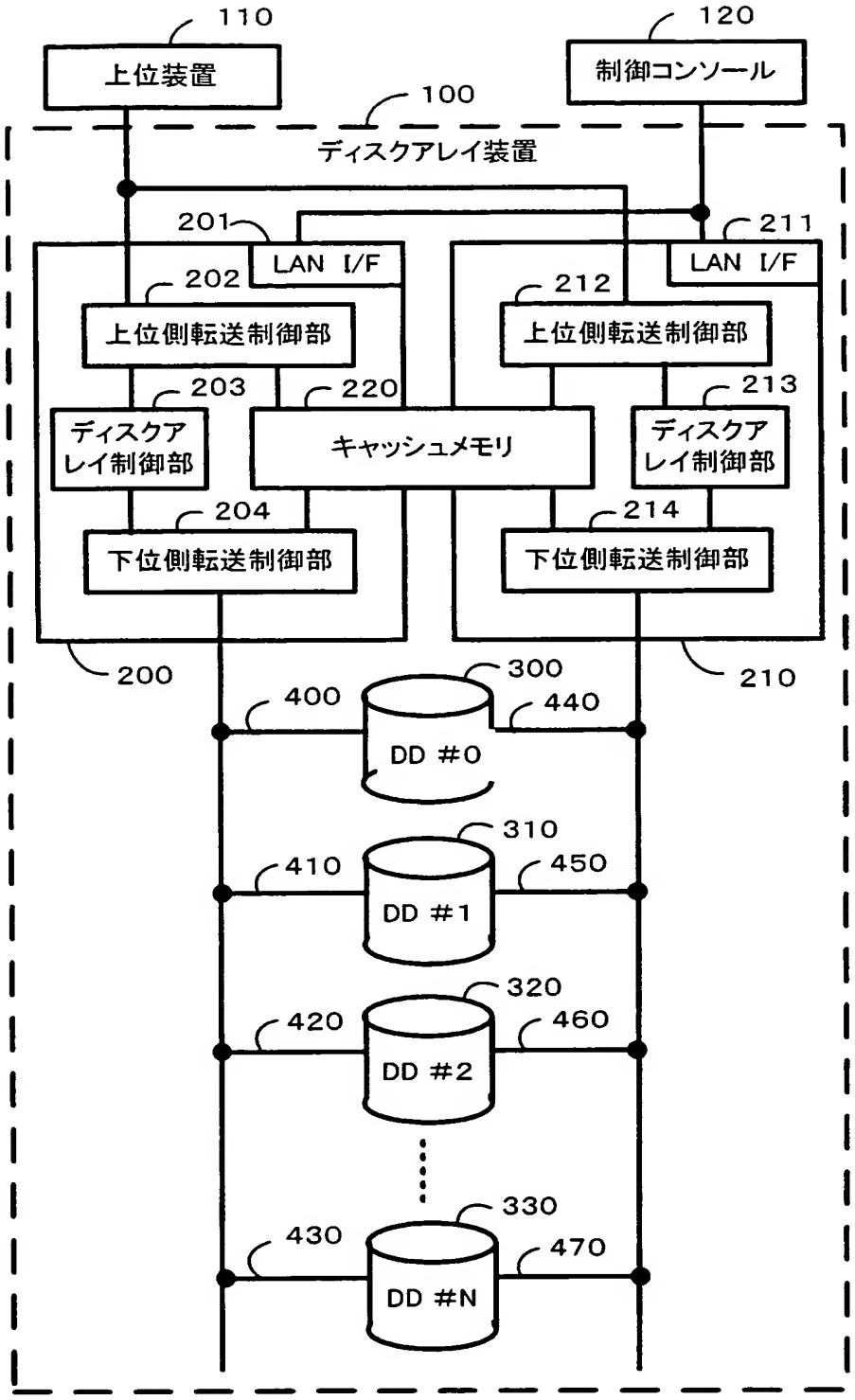
【図１３】パリティグループの構成の変形例を示す模式図。

【符号の説明】

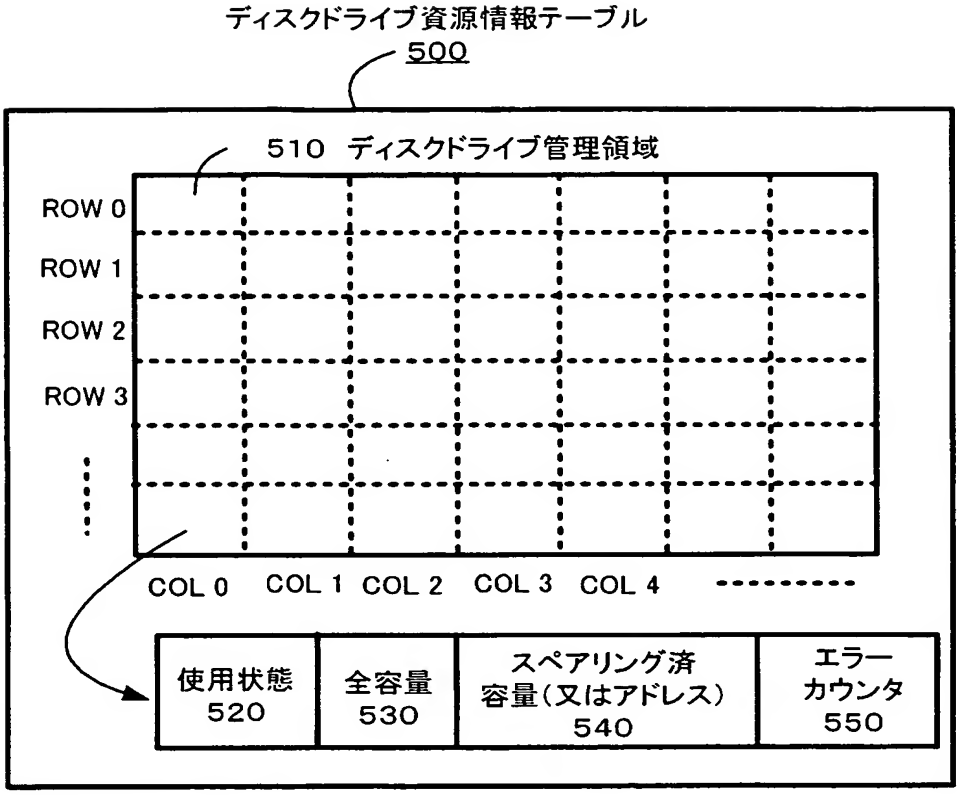
【００６９】

１００…ディスクアレイ装置
１１０…上位装置
１２０…制御コンソール
２００、２１０…コントローラ
２０２，２１２…上位側転送装置
２０３、２１３…ディスクアレイ制御部
２２０…キャッシュメモリ
２０４、２１４…下位側転送装置
３００～３３０…ディスクドライブ#0～#N
５００…ディスクドライブ資源情報テーブル
５１０…ディスクドライブ管理領域
５２０…使用状態
５３０…全容量
５４０…スオペアリング済容量（又はアドレス）
５５０…エラーカウンタ
６００…スペアディスクドライブ資源情報テーブル
６１０…スペアディスクドライブ情報領域
６２０…使用状態
６３０…スペアＲＯＷ
６４０…スペアＣＯＬ
６５０～６７０…使用領域情報[0]～[m]
６７０…使用中フラブ
６８０…開始アドレス
６９０…終了アドレス
７００…コピー元ＲＯＷ
７１０…コピー元ＣＯＬ
８００～８３０…データディスクドライブ#0～#3
８４０…スペアディスクドライブ#A
８５０、１２００、１３００…ディスクドライブ#0の分割コピー元の記憶領域
８６０…ディスクドライブ#0の分割コピー対象外の記憶領域
８７０、１２１０、１３１０…ディスクドライブ#1の分割コピー元の記憶領域
８８０、８９０、１２２０、１２３０、１３４０～１３５０…スペアディスクドライブ#Aの分割コピー先の記憶領域
１３２０…ディスクドライブ#2の分割コピー元の記憶領域

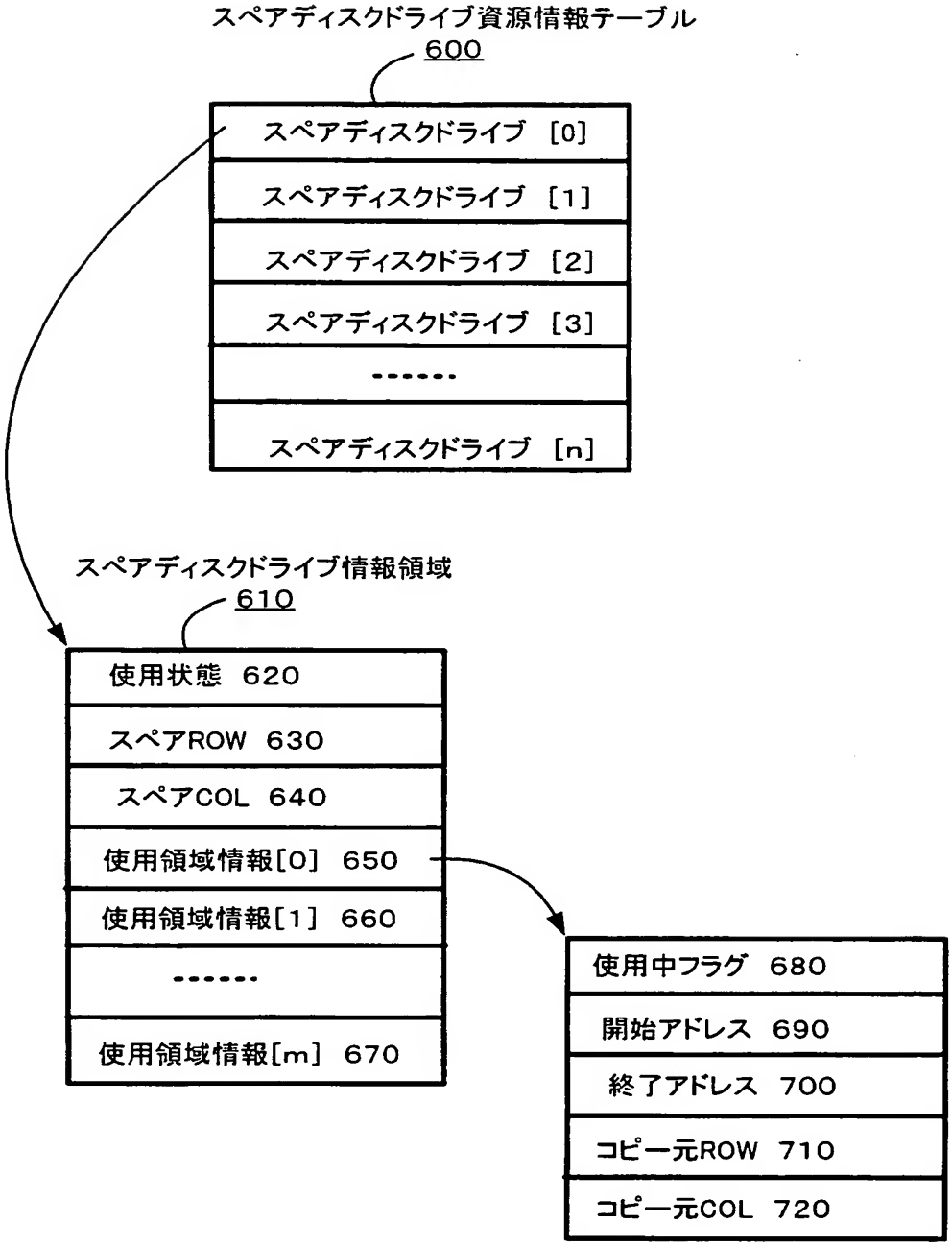
【書類名】図面
【図１】


110
上位装置
120
制御コンソール
100
ディスクアレイ装置
201
LAN I/F
211
LAN I/F
202
上位側転送制御部
212
上位側転送制御部
203
ディスクアレイ制御部
220
キャッシュメモリ
213
ディスクアレイ制御部
204
下位側転送制御部
214
下位側転送制御部
200
210
300
400
440
DD #0
310
410
450
DD #1
320
420
460
DD #2
330
430
470
DD #N

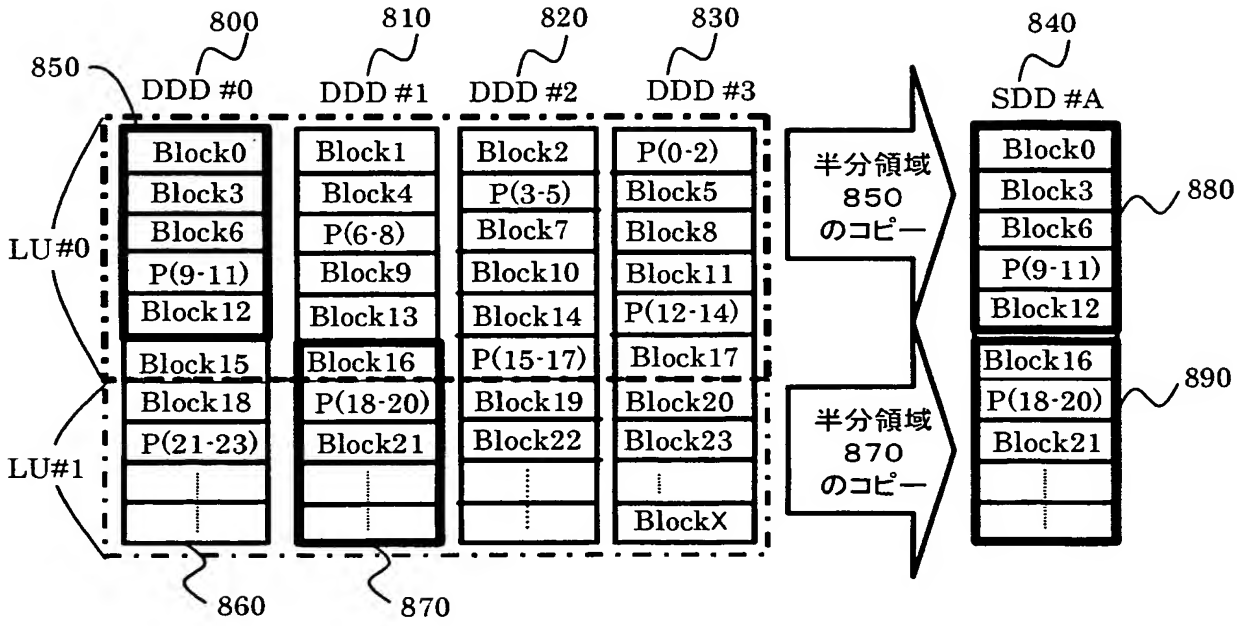
【図２】


ディスクドライブ資源情報テーブル
500
510 ディスクドライブ管理領域
ROW 0
ROW 1
ROW 2
ROW 3
COL 0
COL 1
COL 2
COL 3
COL 4
使用状態
520
全容量
530
スペアリング済
容量(又はアドレス)
540
エラー
カウンタ
550

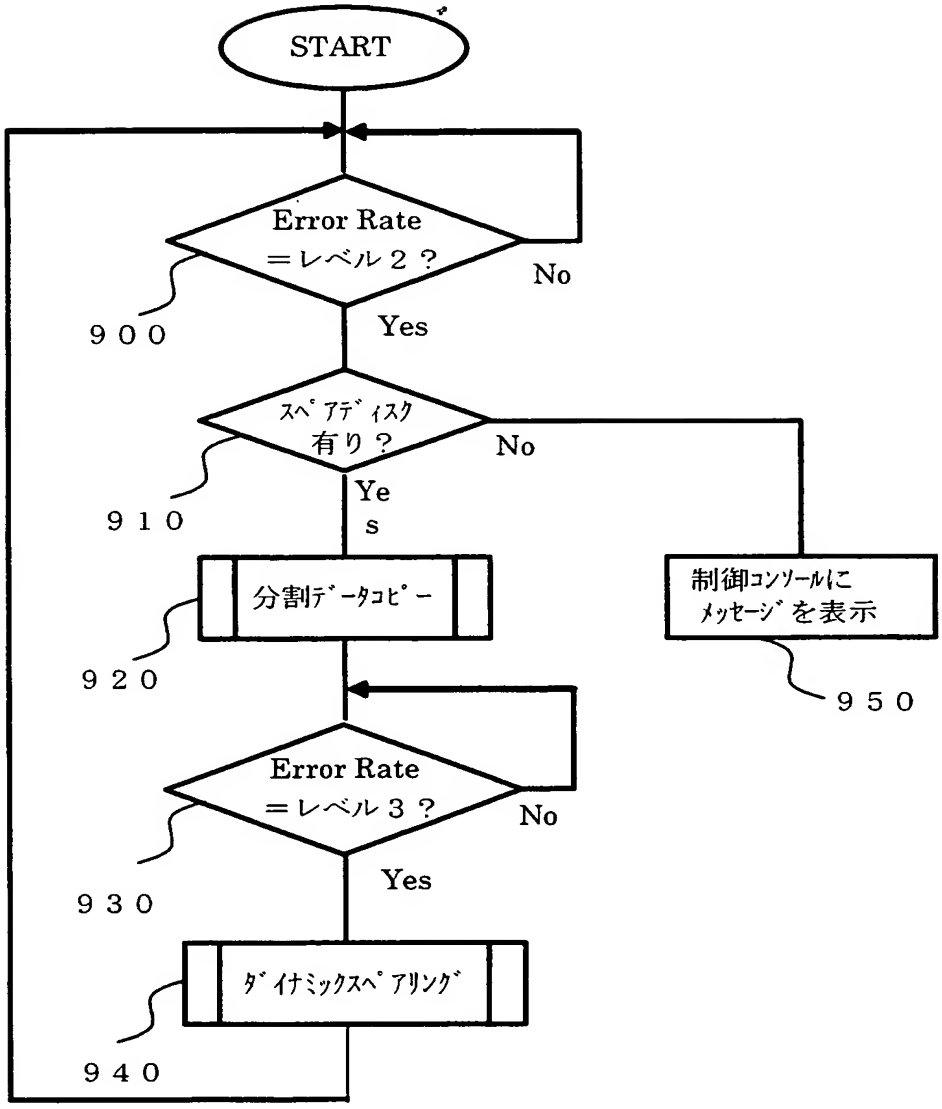
【図３】


スペアディスクドライブ資源情報テーブル
600
スペアディスクドライブ [0]
スペアディスクドライブ [1]
スペアディスクドライブ [2]
スペアディスクドライブ [3]
------
スペアディスクドライブ [n]
スペアディスクドライブ情報領域
610
使用状態 620
スペアROW 630
スペアCOL 640
使用領域情報[0] 650
使用領域情報[1] 660
------
使用領域情報[m] 670
使用中フラグ 680
開始アドレス 690
終了アドレス 700
コピー元ROW 710
コピー元COL 720

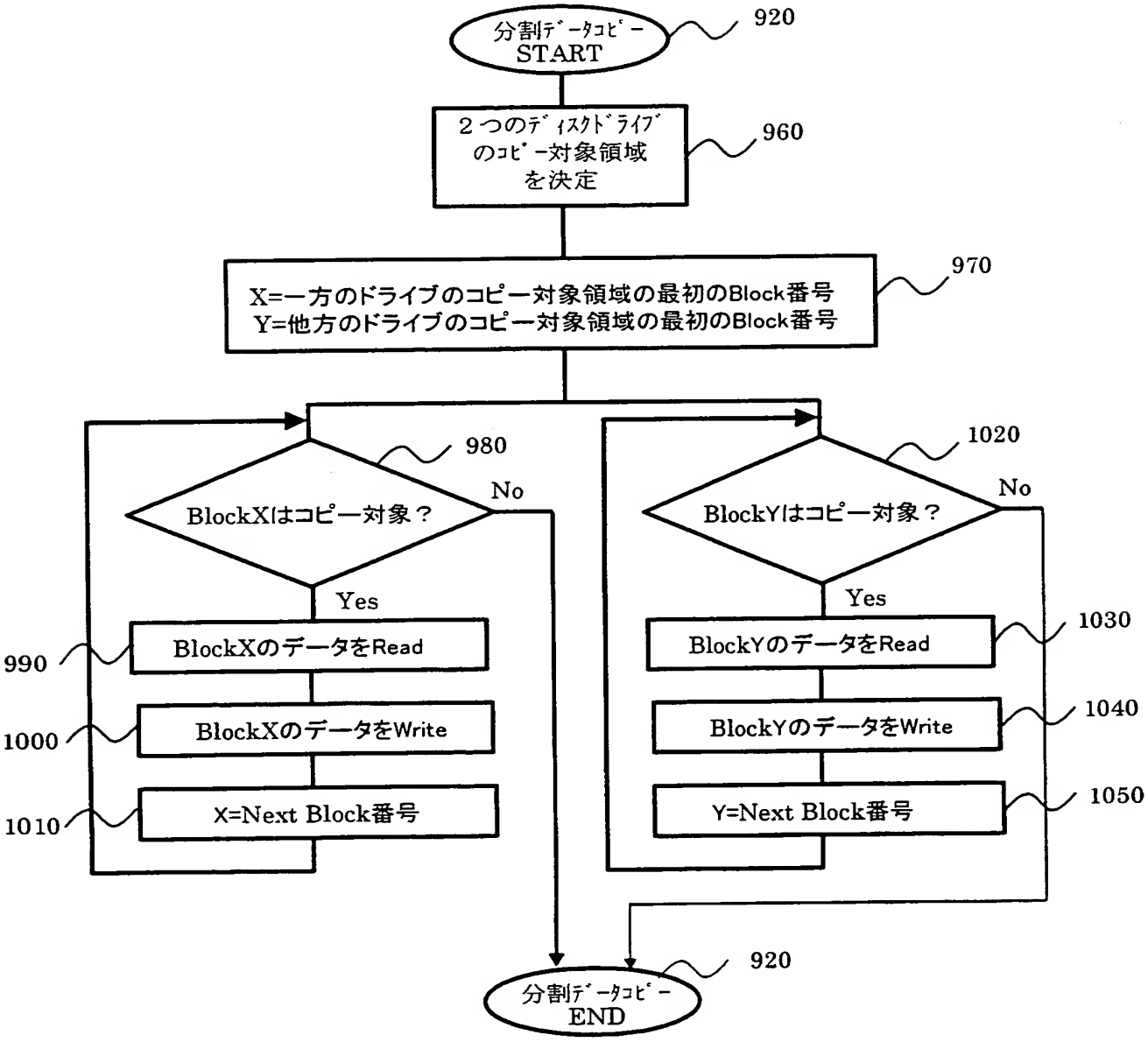
【図４】


800
810
820
830
840
850
DDD #0
DDD #1
DDD #2
DDD #3
SDD #A
LU#0
LU#1
Block0
Block3
Block6
P(9-11)
Block12
Block15
Block18
P(21-23)
Block1
Block4
P(6-8)
Block9
Block13
Block16
P(18-20)
Block21
Block2
P(3-5)
Block7
Block10
Block14
P(15-17)
Block19
Block22
P(0-2)
Block5
Block8
Block11
P(12-14)
Block17
Block20
Block23
BlockX
半分領域
850
のコピー
半分領域
870
のコピー
880
890
860
870

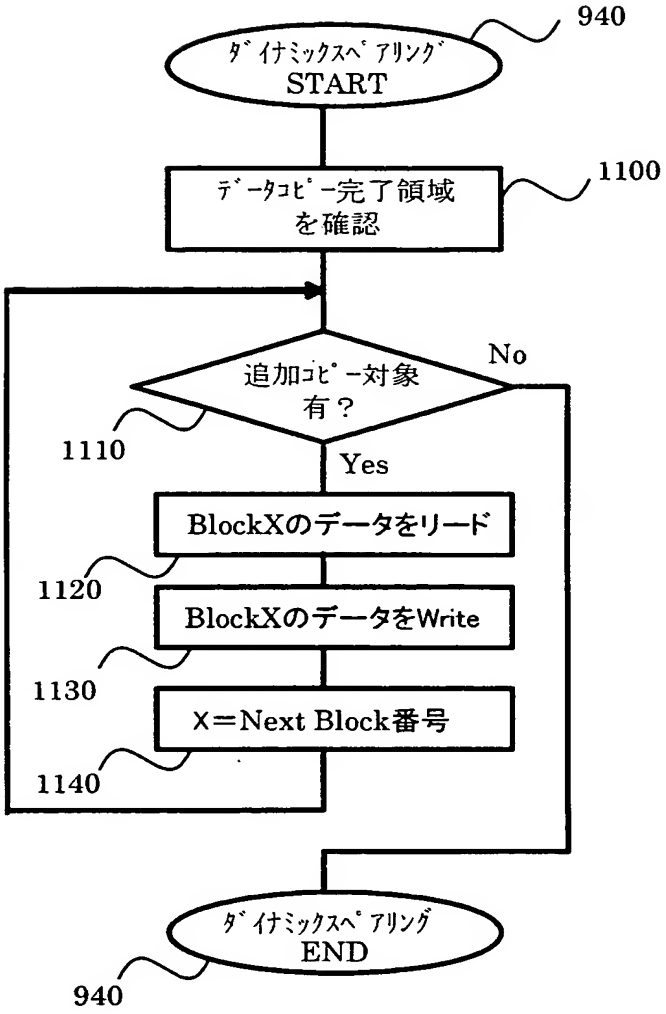
【図５】


START
Error Rate
＝レベル２？
No
９００
Yes
ｽﾍﾟｱﾃﾞｨｽｸ
有り？
No
９１０
Ye
s
分割ﾃﾞｰﾀｺﾋﾟｰ
９２０
制御ｺﾝｿｰﾙに
ﾒｯｾｰｼﾞを表示
９５０
Error Rate
＝レベル３？
No
９３０
Yes
ﾀﾞｲﾅﾐｯｸｽﾍﾟｱﾘﾝｸﾞ
９４０

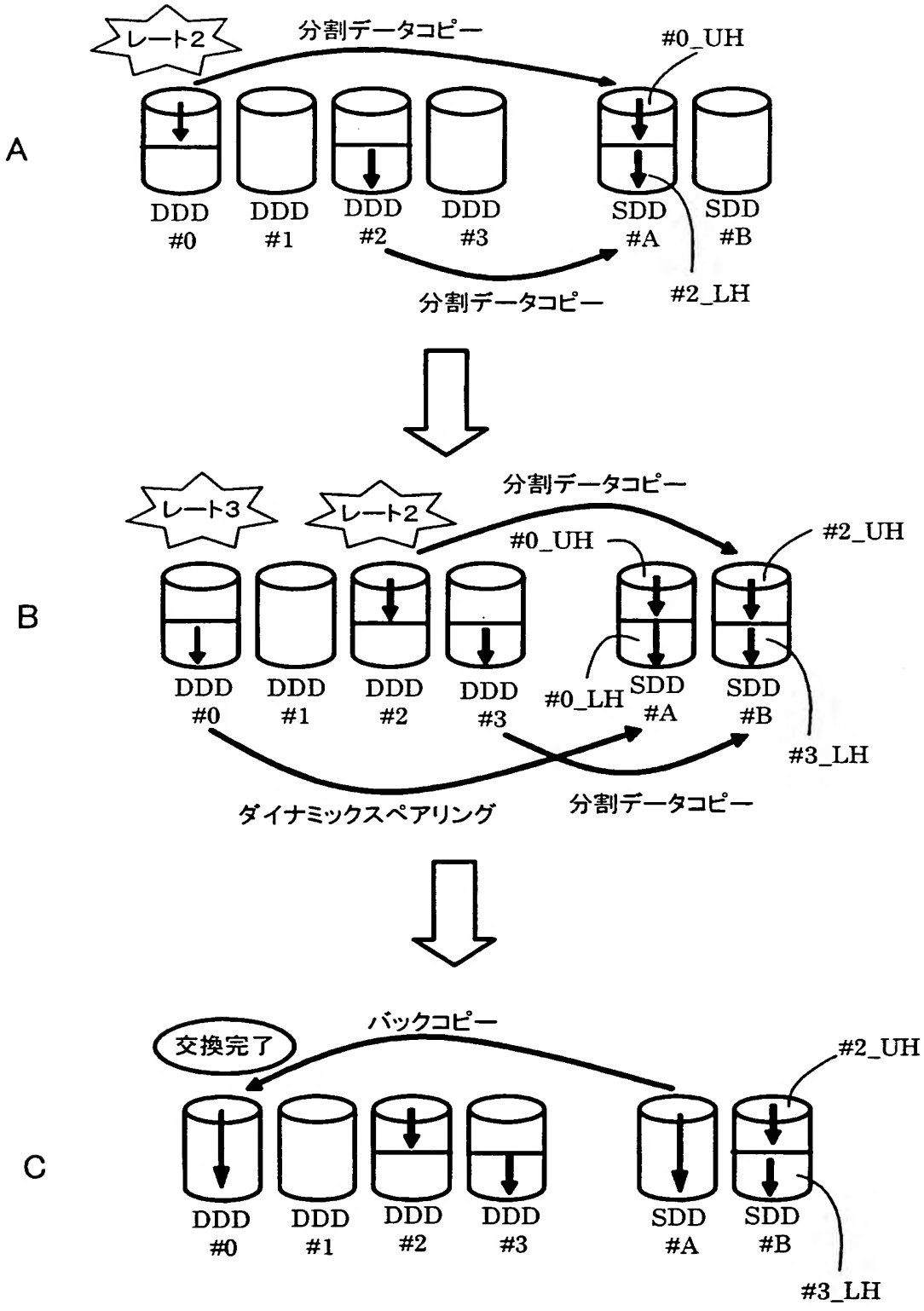
【図６】


分割ﾃﾞｰﾀｺﾋﾟｰ
START
920
２つのﾃﾞｨｽｸﾄﾞﾗｲﾌﾞ
のｺﾋﾟｰ対象領域
を決定
960
X=一方のドライブのコピー対象領域の最初のBlock番号
Y=他方のドライブのコピー対象領域の最初のBlock番号
970
980
BlockXはコピー対象？
No
Yes
990
BlockXのデータをRead
1000
BlockXのデータをWrite
1010
X=Next Block番号
1020
BlockYはコピー対象？
No
Yes
1030
BlockYのデータをRead
1040
BlockYのデータをWrite
1050
Y=Next Block番号
920
分割ﾃﾞｰﾀｺﾋﾟｰ
END

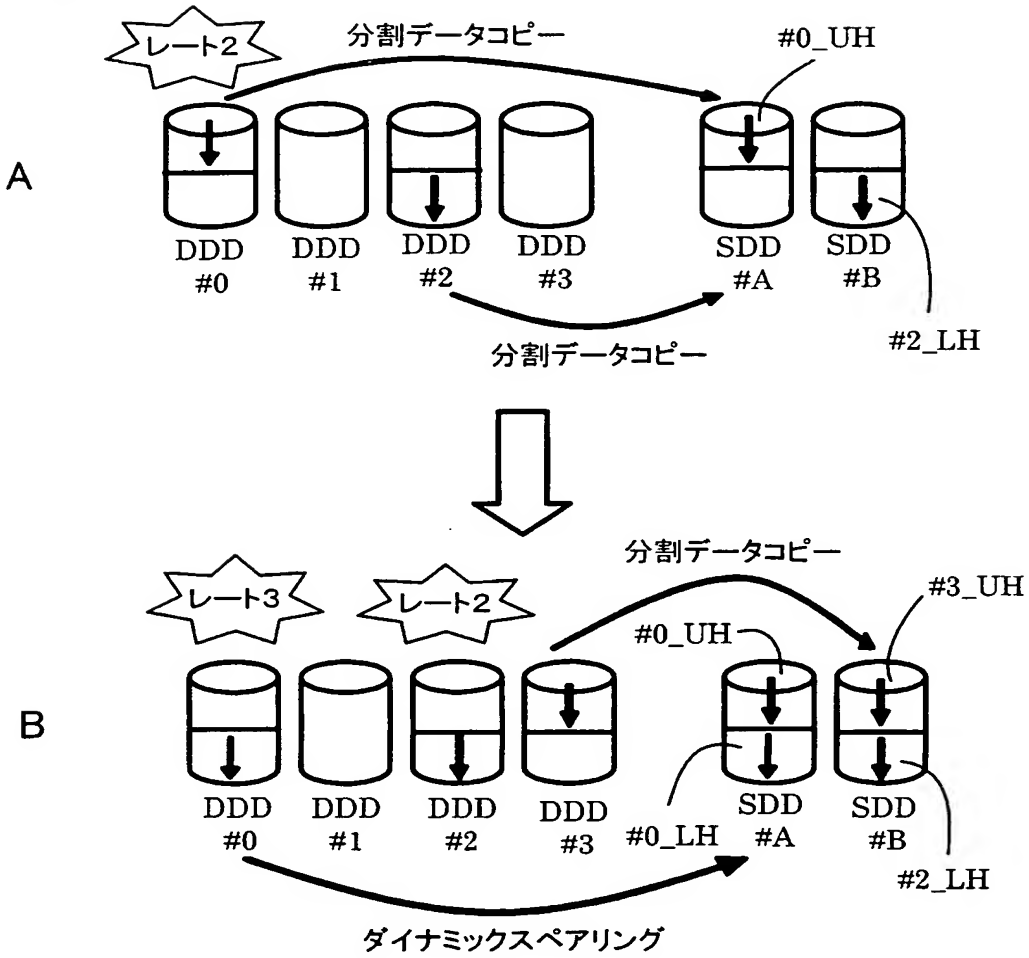
【図７】


940
ﾀﾞｲﾅﾐｯｸｽﾍﾟｱﾘﾝｸﾞ
START
1100
ﾃﾞｰﾀｺﾋﾟｰ完了領域
を確認
No
追加ｺﾋﾟｰ対象
有？
1110
Yes
BlockXのデータをリード
1120
BlockXのデータをWrite
1130
X＝Next Block番号
1140
ﾀﾞｲﾅﾐｯｸｽﾍﾟｱﾘﾝｸﾞ
END
940

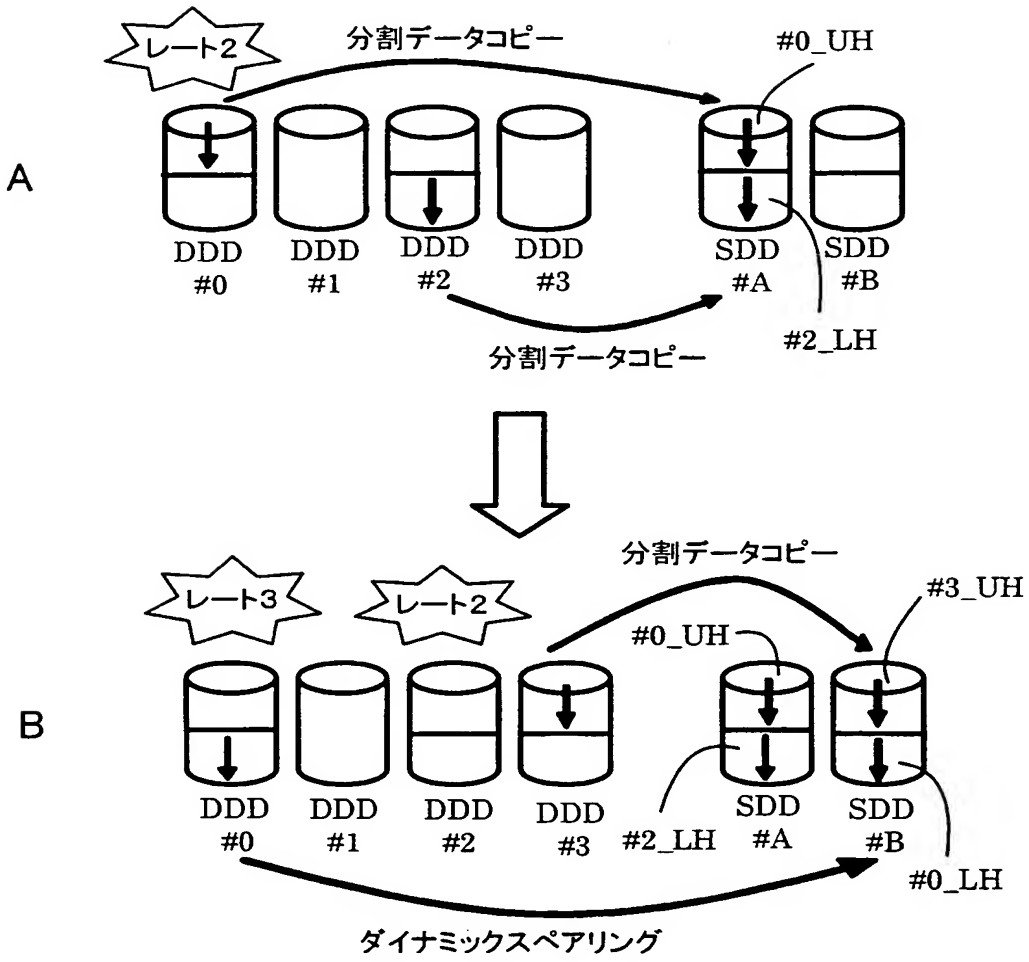
【図８】


A
レート2
分割データコピー
#0_UH
DDD #0
DDD #1
DDD #2
DDD #3
SDD #A
SDD #B
分割データコピー
#2_LH
B
レート3
レート2
分割データコピー
#0_UH
#2_UH
DDD #0
DDD #1
DDD #2
DDD #3
#0_LH
SDD #A
SDD #B
#3_LH
ダイナミックスペアリング
分割データコピー
C
交換完了
バックコピー
#2_UH
DDD #0
DDD #1
DDD #2
DDD #3
SDD #A
SDD #B
#3_LH

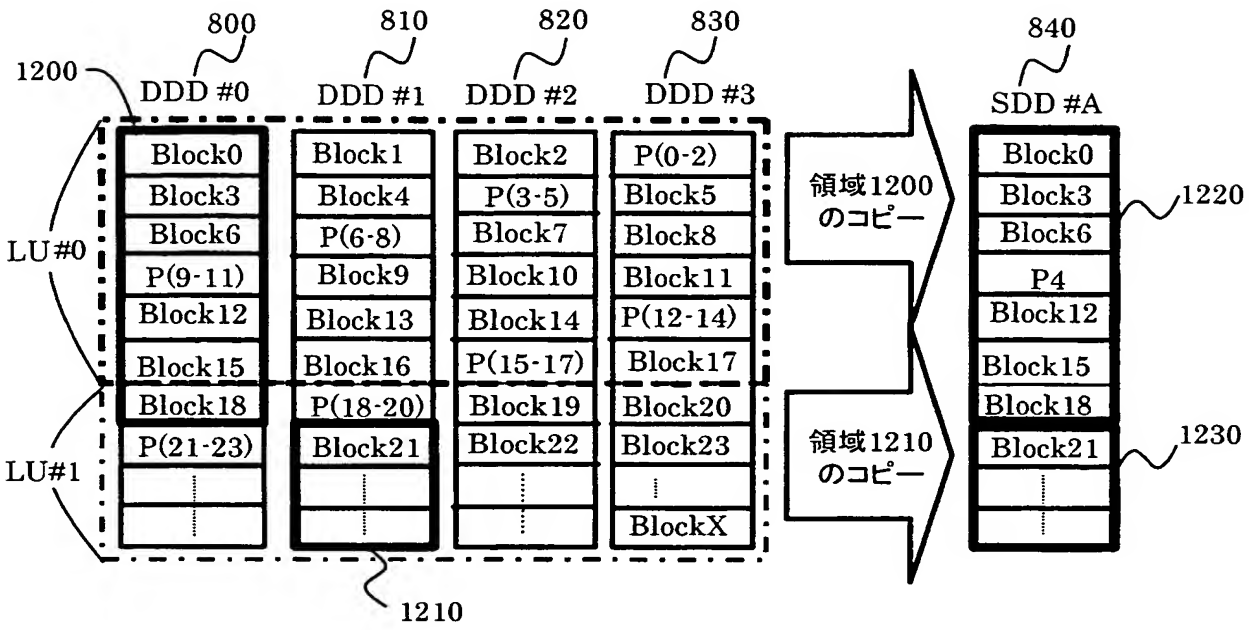
【図９】


A
レート2
分割データコピー
#0_UH
DDD #0
DDD #1
DDD #2
DDD #3
SDD #A
SDD #B
#2_LH
分割データコピー
B
レート3
レート2
分割データコピー
#3_UH
#0_UH
DDD #0
DDD #1
DDD #2
DDD #3
#0_LH
SDD #A
SDD #B
#2_LH
ダイナミックスペアリング

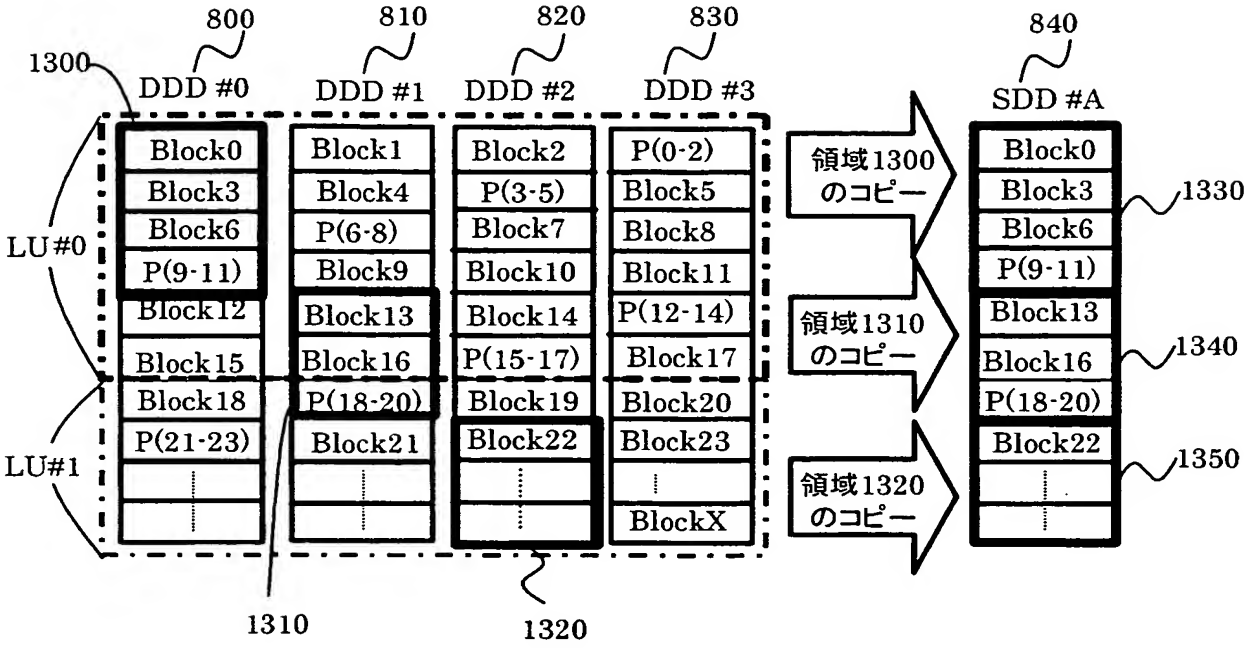
【図１０】


A
レート2
分割データコピー
#0_UH
DDD
#0
DDD
#1
DDD
#2
DDD
#3
SDD
#A
SDD
#B
#2_LH
分割データコピー
B
レート3
レート2
分割データコピー
#3_UH
#0_UH
DDD
#0
DDD
#1
DDD
#2
DDD
#3
#2_LH
SDD
#A
SDD
#B
#0_LH
ダイナミックスペアリング

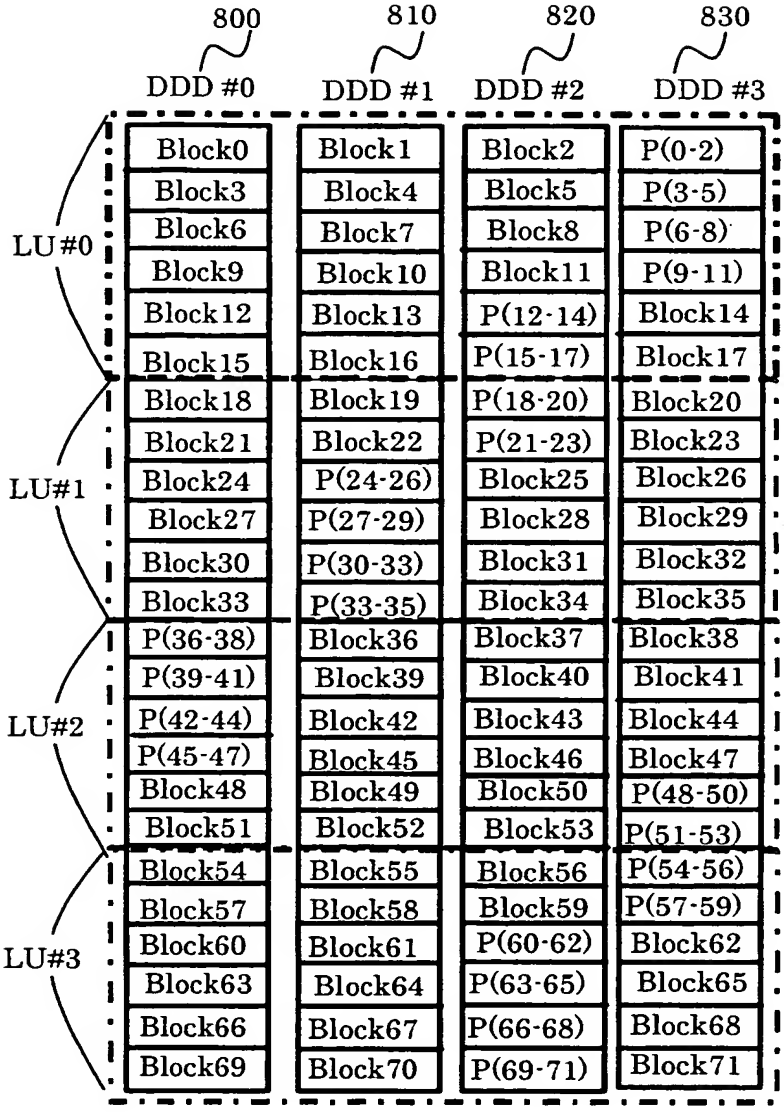
【図１１】


800
810
820
830
840
1200
DDD #0
DDD #1
DDD #2
DDD #3
SDD #A
LU#0
LU#1
Block0
Block3
Block6
P(9-11)
Block12
Block15
Block18
P(21-23)
Block1
Block4
P(6-8)
Block9
Block13
Block16
P(18-20)
Block21
Block2
P(3-5)
Block7
Block10
Block14
P(15-17)
Block19
Block22
P(0-2)
Block5
Block8
Block11
P(12-14)
Block17
Block20
Block23
BlockX
領域1200
のコピー
領域1210
のコピー
Block0
Block3
Block6
P4
Block12
Block15
Block18
Block21
1220
1230
1210

【図１２】


800
810
820
830
840
1300
DDD #0
DDD #1
DDD #2
DDD #3
SDD #A
LU#0
LU#1
Block0
Block3
Block6
P(9-11)
Block12
Block15
Block18
P(21-23)
Block1
Block4
P(6-8)
Block9
Block13
Block16
P(18-20)
Block21
Block2
P(3-5)
Block7
Block10
Block14
Block19
Block22
P(0-2)
Block5
Block8
Block11
P(12-14)
Block17
Block20
Block23
BlockX
領域1300のコピー
領域1310のコピー
領域1320のコピー
1330
1340
1350
1310
1320

【図１３】


800
810
820
830
DDD #0
DDD #1
DDD #2
DDD #3
LU#0
LU#1
LU#2
LU#3
Block0 Block1 Block2 P(0-2)
Block3 Block4 Block5 P(3-5)
Block6 Block7 Block8 P(6-8)
Block9 Block10 Block11 P(9-11)
Block12 Block13 P(12-14) Block14
Block15 Block16 P(15-17) Block17
Block18 Block19 P(18-20) Block20
Block21 Block22 P(21-23) Block23
Block24 P(24-26) Block25 Block26
Block27 P(27-29) Block28 Block29
Block30 P(30-33) Block31 Block32
Block33 P(33-35) Block34 Block35
P(36-38) Block36 Block37 Block38
P(39-41) Block39 Block40 Block41
P(42-44) Block42 Block43 Block44
P(45-47) Block45 Block46 Block47
Block48 Block49 Block50 P(48-50)
Block51 Block52 Block53 P(51-53)
Block54 Block55 Block56 P(54-56)
Block57 Block58 Block59 P(57-59)
Block60 Block61 P(60-62) Block62
Block63 Block64 P(63-65) Block65
Block66 Block67 P(66-68) Block68
Block69 Block70 P(69-71) Block71

【書類名】要約書
【要約】
【課題】　スペアディスクを持つRAID構成のディスクアレイにおいて多重ディスク故障によるデータロストを防止する。
【解決手段】　或るパリティグループ内の個々のディスクドライブ＃０～＃３毎にエラー発生回数から故障発生の可能性を推測し、故障発生の可能性がある程度に高い２つのディスクドライブ＃０、＃２から、データストライプにて互いに異なる２つの半分サイズの記憶領域#0_UHと#2_LHをそれぞれ選び、スペアディスク＃Aへコピーする（分割データコピー）。いずれか一方のディスクドライブ＃０の故障発生の可能性が更に高まると、そのディスクドライブ＃０の残りの半分の記録領域#0_LHをスペアディスク＃Aへコピーする（ダイナミックスペアリング）。分割データコピーにより、ダイナミックスペアリング時間が短縮され、かつ多重ディスク故障によるデータロストの可能性が低減する。
【選択図】　図８

# 認定・付加情報

| | |
|---|---|
| 特許出願の番号 | 特願２００３－３５４５５７ |
| 受付番号 | ５０３０１７０９４００ |
| 書類名 | 特許願 |
| 担当官 | 第七担当上席　　００９６ |
| 作成日 | 平成１５年１０月１６日 |

＜認定情報・付加情報＞

【提出日】　平成15年10月15日

特願２００３－３５４５５７

# 出　願　人　履　歴　情　報

識別番号　［０００００５１０８］

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所